Can we trust "Magnitude-Based Inference"?

Since the times and works of William Sealy Gosset (1876-1937) and Ronald Aylmer Fisher (1890-1962), imperfections of conventional null-hypothesis significance testing and in particular, use of *P*-values to evaluate such testing (invariably referred to as inferential statistics), have been well recognised (Wilkinson, 1999; Wasserstein and Lazar, 2016).

Attempts have been made to identify alternatives. For example, Cohen's effect sizes (Cohen 1988) and region of practical equivalence procedure (ROPE) (Kruschke, 2014). A more recent alternative is magnitude-based inference (MBI) (Hopkins and Baterham, 2016) although unlike others, MBI has created considerable controversy when reporting the results of studies (almost exclusively used in the field of sport and exercise science). Instead of defining research effects as “significant” based on *P*-values (using traditional hypothesis testing), MBI uses terms such as “implementable” and “substantial” based on two constraints called the “risk of harm” and the “chance of benefit”. However, concerns have been raised about the MBI approach. Stanford statistician Kristin Sainani was so concerned about the consequences of using MBI that she wrote a formal analysis of the MBI method. Published in MSSE (Sainani, 2018) her paper showed that, depending on sample size and thresholds for harm/benefit, MBI produces false positive rates that can be two to six times greater than those using traditional hypothesis testing. A finding, she claims, that makes MBI less reliable.

Loosening or lowering the standard of evidence (increasing the false positive rates) has important consequences. Of course, we recognise the reciprocal nature of error reduction i.e. by increasing Type I error rates (false-positives), we automatically reduce Type II (false-negative) rates and *vice versa*. Ethically, declaring that many interventions work when they do not, is unacceptable to editors and practitioners alike. For example, an athlete could adopt several interventions, many of which might not work. This is wasteful in time, cost and energy. Sainani provided several examples, including those from published papers, where MBI indicated either implementable or substantial interventions that were not statistically significant.

As such, Medicine and Science in Sport and Exercise (MSSE) has decided not to accept for publication papers that use MBI. After consultation with the journal’s Editorial Board, the Editor in Chief (Bruce Gladden) has decided not to allow MBI until a properly vetted account of the method has been published in a recognized statistics journal (and he recommended that researchers should not use MBI until that occurs).

There is another concern with MBI and traditional hypothesis testing used in sport and exercise research. All MBI, and the majority of traditional hypothesis-testing inferences, are based on confidence intervals that assume data are symmetric and normally distributed (e.g., all confidence intervals used in MBI are calculated using an appropriate *t*-value, say at the 95% probability, multiplied by the standard error of the mean). The majority of data reported in sport and exercise science are measured on the ratio rather than the interval scale (e.g., maximum oxygen uptake, strength and speed). Because ratio data cannot be negative (such data are bounded to the left by zero but invariably unbounded to the right), ratio data tend to be positively

skewed and, as such, will not be symmetric (i.e., using symmetric confidence intervals is likely to be misleading). This positive skew characteristic in such ratio data can be overcome by taking logarithms (see Nevill 1997). Neither supporters of MBI nor its opponents, including Sainani (2018), appear to be concerned about whether the data they are considering are, or are not, symmetric and satisfy normal-distribution assumptions, assumptions that need to be assumed when deciding whether or not an effect is likely to be real.

In 2014, the Journal of Sports Sciences (Winter et al., 2014) outlined its revised policy on how to report statistical inference. Manuscripts would not be accepted for publication if analyses were evaluated solely based on *P*-values. Authors were encouraged to provide supplementary statistics such as effect sizes and confidence intervals (usually the 90% or 95%) that would provide readers with additional information to assess the benefit of an intervention. This policy reflects a cautious approach that erred on the side of evolution rather than revolution (i.e., *P*-values were not barred completely and specific alternatives were not stated as preferences).

The debate as to whether MBI is a suitable alternative to *P*-values is likely to continue for some time. However, following the Journal's traditionally conservative approach, the Editorial Board is unanimous in its view that until MBI receives formal endorsement from academic and medical statisticians alike (e.g., an endorsement published in reputable journals such as the Royal Statistical Society and the BMJ respectively), MBI should be used with caution. When choosing which methods of inference to report, we prefer authors to adopt the recommendations outlined in the editorial written by Winter et al. (2014).

Alan Nevill, Editorial Board, Edward Winter,

References

Cohen, J. (1988). Statistical power analysis for social sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hopkins, W. G., & Batterham, A. M. (2016). Error rates, decisive outcomes and publication bias with several inferential methods. *Sports Medicine*, *46*(10), 1563-1573.

Kruschke, J. (2014). Doing bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd ed.). London: Academic Press.

Nevill, A.M. "Why the analysis of performance variables recorded on a ratio scale will invariably benefit from a log transformation". *Journal of Sports Sciences*, **15**, 457-458, 1997.

Sainani K.L. The Problem with" Magnitude-Based Inference". Medicine and Science in Sports and Exercise (2018). DOI: 10.1249/MSS.0000000000001645.Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. The American Statistician, 70, 129-133.

Wilkinson, L. & Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-564.

Winter, E.M., Abt, G.A. and Nevill, A.M. (2014). Metrics of meaningfulness as opposed to sleights of significance. Journal of Sports Sciences, 32, 901-902.